An Introduction to the DIMA and IAM Groups

Prof. Dr. Volker Markl

Chair of the Database Systems and Information Management (DIMA) Group at TU Berlin
Director of the Intelligent Analytics for Massive Data (IAM) Group at DFKI

volker.markl@tu-berlin.de, http://www.user.tu-berlin.de/marklv
Technische Universität Berlin - Faculty IV EECS

One of Germany’s largest and internationally renowned institutes of technology, including:

- 7 Faculties with 366 Professors and 2,650+ Researchers (2019)
- 35,500+ Students (2019)
- 40+ Bachelor’s and 70+ Master’s Programmes
- About 180M € in Third Party Funding (2018)

Innovative Alliances and Prestigious, Award-Winning Research:

- 6,100+ Students (30% International)
- 70+ Professors (including Dual, Guest and Junior Professors)
- Over 500 Scientific Staff Members
- 39*M € in External Funding (2018)

The collective Faculty IV research expertise is organized across six labs: (1) Cognitive Systems, (2) Cyber-Physical Systems, (3) **Data Analytics and Cloud**, Future Internet Media and Technology, Integration Technology and Photonics, and Sustainable Energy Systems and Mobility.
German Research Center for Artificial Intelligence

» Locations
- Robotics Innovation Center
- Cyber Physical Systems
- Oldenburg
- Bremen
- Berlin
- Osnabrück
- Saarbrücken
- Kaiserslautern
- Marine Perception
- Plan-Based Robot Control
- Smart Enterprise Engineering
- Agents and Simulated Reality
- Innovative Retail Laboratory
- Intelligent User Interfaces
- Institute for Information Systems
- Multilingual Technologies
- Smart Service Engineering
- Embedded Intelligence
- Augmented Vision
- Innovative Factory Systems
- Intelligent Networks
- Smart Data and Knowledge Services
- Intelligent Analytics for Massive Data
- Language Technology
- Interactive Textiles
- Educational Technology Lab

» Shareholders
- 19 Research Groups
- 255 Active Projects
- 1000+ Employees (756 FTE)
- 80,2 M € 2018
### About Volker Markl

**Publications**: 225+, **Patents**: 18, **Mentor/Co-founder of Startups**: 7

**Funding Grants**: 65+ (Since 2008): 30+ publicly funded, 30+ industry funded

**Awards**: SIGMOD 2020 Best Paper Award, BTW Data Science Challenge 2019 Winner, European Information Society and Technology Prize, VLDB and ECBT Best Paper Awards, ACM SIGMOD Research Highlight Award, IBM Faculty Awards, and named one of Germany’s leading minds in IT (*Deutschlands Digitale Köpfe*) in 2015

<table>
<thead>
<tr>
<th>Position</th>
<th>Organization</th>
<th>Funded by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Professor, Chair</td>
<td>Database Systems and Information Management (DIMA) Group, Technische Universität Berlin</td>
<td></td>
</tr>
<tr>
<td>Director</td>
<td>Intelligent Analytics for Massive Data Research Group (IAM), Funded by the German Research Center for Artificial Intelligence (DFKI)</td>
<td></td>
</tr>
<tr>
<td>Director</td>
<td>Berlin Institute for the Foundations of Learning and Data (BIFOLD) Funded by the German Federal Ministry of Education and Research (BMBF)</td>
<td></td>
</tr>
<tr>
<td>President</td>
<td>VLDB Endowment</td>
<td></td>
</tr>
<tr>
<td>Speaker</td>
<td>Forum Digital Technologies (FDT), Funded by the German Federal Ministry for Economic Affairs and Energy (BMWi)</td>
<td></td>
</tr>
<tr>
<td>Speaker</td>
<td>Data Analytics Laboratory, Fak. IV School of Electrical Engineering and Computer Science (EECS), Technische Universität Berlin</td>
<td></td>
</tr>
<tr>
<td>Founding Member</td>
<td>Big Data Value Association (BDVA)</td>
<td></td>
</tr>
</tbody>
</table>
Research Vision

Our goal is to help bridge the data analytics talent gap through research and the development (R&D) of novel technologies, systems, and methods along the entire data value chain, including data acquisition, information extraction and integration, model building, and interactive exploration.

Our R&D activities are centered on the delivery of cutting-edge systems, technologies, and tools for the storage, processing, and low latency management of massive amounts of heterogeneous data that address challenges along the entire data value chain, including

- **Systems: Scalability and Performance**
- **Systems: Programmability and Ease of Use**
- **Development of Scalable Data Analytics**
- **Exploitation of Modern Hardware to Accelerate Processing**
- **Benchmarking and Performance Analysis**

### Total Staff (TUB/DFKI)
- Professors: 02
- Senior Staff: 14

© Volker Markl
Shaping Global Initiatives

<table>
<thead>
<tr>
<th>Initiative</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIFOLD</td>
<td>The Berlin Institute for the Foundations of Learning and Data aims to conduct research into the scientific foundations of Big Data and Machine Learning, to advance AI application development, and greatly increase the impact to society, the economy, and science.</td>
<td><a href="https://bifold.berlin">https://bifold.berlin</a></td>
</tr>
<tr>
<td>Very Large Data Base (VLDB) Endowment</td>
<td>Europe is home to over 55 Centers of Excelldence that perform cutting edge research and drive the development and evolution of all aspects of Big Data. The vision of the network is to landscape and connect Europe’s Big Data Competence.</td>
<td><a href="http://www.vldb.org">http://www.vldb.org</a></td>
</tr>
<tr>
<td>Plattform Lernende Systeme</td>
<td>The Plattform Lernende Systeme unites experts from science, industry, &amp; society in technological, economic and social working groups (WGs) to discuss challenges in the design of learning systems that address human needs.</td>
<td><a href="https://www.plattform-lernende-systeme.de/home-en.html">https://www.plattform-lernende-systeme.de/home-en.html</a></td>
</tr>
<tr>
<td>Forum Digital Technologies</td>
<td>The Forum Digital Technologies is a networking platform and a demonstrator hub for digital technologies and innovations made in Germany. Its Berlin based showroom showcases German research in this field. Through (inter)national roadshows, the forum fosters the visibility of German digital innovations and the establishment of new partnerships.</td>
<td><a href="https://www.digitale-technologien.de">https://www.digitale-technologien.de</a></td>
</tr>
<tr>
<td>Einstein Center Digital Future (ECDF)</td>
<td>The Einstein Center Digital Future (ECDF) is an interuniversity nucleus for research on the digitization of society. It aims to foster innovative, cutting-edge interdisciplinary research, and provide outstanding training for talented young scholars.</td>
<td><a href="https://www.digital-future.berlin">https://www.digital-future.berlin</a></td>
</tr>
</tbody>
</table>

Very Large Data Base (VLDB) Endowment is a non-profit organisation incorporated in the United States for the sole purpose of promoting and exchanging scholarly work in databases and related fields throughout the world.

Europe is home to over 55 Centers of Excellence that perform cutting edge research and drive the development and evolution of all aspects of Big Data. The vision of the network is to landscape and connect Europe’s Big Data Competence.

The Plattform Lernende Systeme unites experts from science, industry, & society in technological, economic and social working groups (WGs) to discuss challenges in the design of learning systems that address human needs.

The Forum Digital Technologies is a networking platform and a demonstrator hub for digital technologies and innovations made in Germany. Its Berlin based showroom showcases German research in this field. Through (inter)national roadshows, the forum fosters the visibility of German digital innovations and the establishment of new partnerships.

The Einstein Center Digital Future (ECDF) is an interuniversity nucleus for research on the digitization of society. It aims to foster innovative, cutting-edge interdisciplinary research, and provide outstanding training for talented young scholars.
BIFOLD will pursue the following strategic priorities in line with the German National AI Strategy:

**Research**: Conduct high-impact foundational research in the fields of Big Data, Machine Learning and their intersection, to advance the state-of-the-art in Big Data and Machine Learning methods and technologies as well as attract the world’s best scientists to Germany.

**Innovation**: Prototype AI technologies, Big Data systems, Data Science tools, Machine Learning algorithms, and support knowledge and technology exchange, to empower innovation in the sciences, humanities, and companies, particularly, startups.

**Education**: Prepare the next generation of experts in Big Data and Machine Learning, for future academic or industrial careers.

Involved Institutions:
Big Data Analytics = Machine Learning + Data Management

**Machine Learning**
- Feature Engineering
- Representation Algorithms (e.g., SVMs, GPs)
- Control Flows
- Linear Algebra
- Monte Carlo
- Error Estimation
- Regression

**Data Management**
- Data Warehouse / OLAP
- NF2 / XQuery
- Relational Algebra / SQL
- Hardware Adaptation
- Fault Tolerance
- Resource Management
- Data Flows
- Hardware Adaptation
- Indexing

**Technology**
- Declarative Languages
- Automatic Adaptation
- Scalable Processing
- Parallelization
- Compilers
- Memory Management
- Memory Hierarchies
- Data Analysis Languages
- Query Optimization
- Data Flows
- Indexing

**Think ML algorithms in a scalable way**

**Process iterative algorithms in a scalable way**

**End Goal:**
**Data analysis without systems programming!**
Aspects of Data Science
(e.g., logistics, medicine, physics, mechanical engineering, energy)

Mathematical Programming
Linear Algebra
Stochastic Gradient Descent
Explainable AI
Sparse Data
Regression
Monte Carlo
Statistics
Sketches
Hashing
Convergence
Logics and Planning
Deep Learning
Curse of Dimensionality

Relational Algebra / SQL
Data Warehouse / OLAP
HDFS / KVS
Information Extraction
Hardware Adaption
Information Integration
Storage Management
Parallelization
Scalability / Latency
Reproducibility
Data Analysis Languages
Query Optimization
Indexing
Architectures
Responsible Data Management

Conclusion: Data scientists must be talented all around.
Forum Digital Technologies

Business Associations & Initiatives
- bitkom
- VOICE
- bvmj
- eco
- ZVCI: VDE VDA
- BMW
- Branchenverbände
- INDUSTRIE 4.0
- Lernende Systeme
- INDUSTRIAL DATA SPACE
- Smart Data Innovation Lab

Politics & Society
- Bundesministerium für Wirtschaft und Energie
- Bundesministerium für Bildung und Forschung

Research Projects & Programs
- Smart Data
- BIG DATA
- Fraunhofer BIG DATA

Meeting Point and Showroom

Engage SMEs

Foster International Exchange
Towards Open Environments

An Ecosystem for AI and a Platform for the IoT

Agora
An Open Ecosystem for Data Science & Artificial Intelligence


NebulaStream
Data and Application Management for the Internet of Things


© Volker Markl
An asset-centric ecosystem, where everyone can offer and access all kinds of data-related assets – and – combine them to ease the development of novel applications, regardless of expertise.

Key Features

- Outlining a data infrastructure architecture:
  - Combining data-related assets to novel applications
  - Automating and optimizing the application execution

- Highlighting research challenges:
  - Asset Marketplace
  - Asset Standardization and Certification
  - Operations Manager

Key Publication

Jonas Traub, Jorge-Arnulfo Quiané-Ruiz, Zoi Kaoudi, and Volker Markl

Agora Operations Manager

1. Register compute nodes at marketplaces
2. Synchronize node catalogs
3. Compose asset-based applications
4. Send apps to the operations manager for deployment
5. Match applications with compute nodes
6. Deploy processing pipelines
7. Retrieve application output
Agora

Selected Research Challenges

Asset Marketplaces
- Asset Sharing
- Asset Discovery
- Asset Search
- Marketplace Regulations

Asset Standardization and Certification
- Unified Asset Specification
- Hierarchies and Compatibilities of Assets
- Trustful Certification of Assets
- Combining Generality and Simplicity

Operations Manager
- Matching Applications and Compute Nodes
- Asset Usage Tracking
- Query Optimization
- Privacy and Security
NebulaStream addresses the heterogeneity and distribution of both compute and data, supports diverse data and programming models going beyond relational algebra, copes with potentially unreliable communication, and enables constant evolution under continuous operation.

Key Features

- Bounded and unbounded data streams over millions of sensors
- Concurrent execution of thousands of queries
- Stateful and fault-tolerant processing from sensors to cloud
- Hardware-tailored code generation for maximal efficiency among highly heterogeneous environments
- By-design support for mobility
- Unified support for relational, linear, and graph algebra in a single engine

Key Publication

Steffen Zeuch, Ankit Chaudhary, Bonaventura Del Monte, Haralambos Gavriilidis, Dimitrios Giouroukis, Philipp M. Grulich, Sebastian Breß, Jonas Traub, and Volker Markl

NebulaStream: An Application Scenario

Diversity in Programming and Management Environments

Constant Evolution under Continuous Operation

Heterogeneity, Distribution, and Volume of both Data at Rest and Data in Motion

Legend:
- Control Flow
- Base Station
- Data Flow
- Processing Node
- Vehicle/Sensor Node
- Disconnected Vehicle

Spontaneous, Potentially Unreliable Connectivity between Data and Compute

Feedback Loops between the Real World and Digital Twins

Heterogeneity, Distribution, and Volume of Compute
NebulaStream: Initial Experiments

Saving network traffic with in-network (pre-) processing

Increasing efficiency on low-end nodes with query compilation

Scaling to large numbers of nodes with decentralized joins

Improving fault tolerance for an unreliable environment

NebulaStream: Initial Experiments

Saving network traffic with in-network (pre-) processing

Increasing efficiency on low-end nodes with query compilation

Scaling to large numbers of nodes with decentralized joins

Improving fault tolerance for an unreliable environment
Key Successes

Apache Flink

Emma

Peel

Hawk

Myriad
Apache Flink® is an open-source stream processing framework for distributed, high-performing, always-available, and accurate data streaming applications.

Key Features

- Bounded and unbounded data
- Event time semantics
- Stateful and fault-tolerant
- Running on thousands of nodes with very good throughput and latency
- Exactly-once semantics for stateful computations
- Flexible windowing based on time, count, or sessions in addition to data-driven windows

Key Publications

P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas:

A. Alexandrov, V. Markl, et al:
case class Path (from: Long, to: Long)
val tc = edges.iterate(10) {
  paths: DataSet[Path] =>
  val next = paths
    .join(edges)
    .where("to")
    .equalTo("from") {
      (path, edge) =>
        Path(path.from, edge.to)
    }.
  .union(paths)
  .distinct()
  next
}

D. Battré, S. Ewen, F. Hueske, O. Kao, V. Markl, and D. Warneke:
The Effect of Optimization

- Execute a sample on a laptop
- Execute Plan A
- Execute Plan B
- Execute Plan C
- Hash vs. Sort
- Partition vs. Broadcast
- Caching
- Reusing Partition / Sort
- Run large files on a cluster
- Run a month later, after the data has evolved
Do you want to hand-tune that?

Processing Iterative Data Analysis Programs

Pushing work “out of the loop”

Caching Loop-invariant Data

Maintain state as index

Factorizing a matrix with 28 billion ratings for recommendations

(Scale of Netflix or Spotify)


S. Ewen, S. Schelter, K. Tzoumas, D. Warneke, and V. Markl:

S. Ewen, K. Tzoumas, M. Kaufmann, and V. Markl:
Streaming: Some Benchmark Results

Initially performed by Yahoo! Engineering

[..]Storm 0.10.0, 0.11.0-SNAPSHOT and Flink 0.10.1 show sub-second latencies at relatively high throughputs[..]. Spark streaming 1.5.1 supports high throughputs, but at a relatively higher latency.
Apache Flink

Flink Community

- 26,500+ Meetup Members Worldwide
- 670+ Open Source Contributors/Developers
- 46 Meetup Groups Worldwide

Flink Contributors

- 18 Countries that Regularly Hold Meetups
- 49+ Companies using Apache Flink
- Startup data Artisans (now Ververica)

https://www.meetup.com/topics/apache-flink/
https://flink.apache.org/poweredby.html
https://github.com/apache/flink

Last updated: June 2020
Emma is a quotation-based Scala DSL that enables holistic optimizations of data flow programs for scalable data analysis on Apache Flink and Spark.

Key Features

- Declarative DSL based on Scala-native constructs that increases developer productivity.
- Enables holistic optimizations that integrate database and compiler techniques and include control flows.
- Targets Apache Flink and Spark for execution and is easily extendable to new systems.

Key Publications


Hawk is a hardware-tailored code generator, which enables database queries to run efficiently on various heterogeneous processors without any manual tuning.

**Key Features**

- Generates custom code for database queries for heterogeneous processors, such as CPUs, GPUs, and MICs.
- Automatically rewrites its generated code until queries run with peak performance.
- Introduces novel query execution strategies for GPUs.

**Key Publications**

*Henning Funke, Sebastian Breß, Stefan Noll, Volker Markl, and Jens Teubner:*

*Sebastian Breß, Bastian Köcher, Henning Funke, Tilmann Rabl, and Volker Markl:*

[http://cogadb.dfki.de](http://cogadb.dfki.de)
The Myriad Toolkit facilitates the specification of scalable data generation programs with complex statistical constraints via a special XML data generator prototyping language.

Key Features

- Simplifies generating large, synthetic datasets with prescribed schemas and a set of statistical constraints. Provides a fast and easy way to develop data generators that can produce statistically dependent data in parallel on a set of independently running nodes.

- A challenging yet increasingly important task, especially in the context of benchmarking and testing systems for web-scale data management or parallel RDBMS (e.g., Hadoop, DB2).

Key Publication

Alexander Alexandrov, Berni Schiefer, John Poelman, Stephan Ewen, Thomas Bodner, and Volker Markl:
Peel is a framework that helps define, execute, analyze, and share experiments for distributed systems and algorithms.

Key Features

- Enables the transparent specification of benchmarking workloads and system configuration parameters.
- Orchestrates the systems involved and automatically runs experiments and collects all associated experiment logs.
- Currently, supports Apache HDFS, Hadoop, Flink, and Spark and can easily be extended to include additional systems.

Key Publication

Christoph Boden, Alexander Alexandrov, Andreas Kunft, Tilmann Rabl, and Volker Markl.


http://peel-framework.org/
Current Projects

**FogGuru** contributes to the rapidly emerging domain of fog computing with technologies for managing application resources, middleware for easing the development of innovative applications, and application blueprints to demonstrate how these technologies can provide tangible benefits to the European citizens. *Our group* will train talented early-stage researchers with an innovative and intersectoral research program to constitute the next generation of European Cloud and Fog computing experts. [http://www.fogguru.eu](http://www.fogguru.eu)

**E2Data** will provide a new big data software paradigm to achieve the maximum resource utilization for heterogeneous cloud deployments without affecting current big data programming norms. *Our group* provides solutions in the core of the big data stack. We lead the effort in providing the appropriate software components that will enable key functions in Flink to execute on heterogeneous resources. [https://e2data.eu](https://e2data.eu)

**Hawk** will develop concepts to enable database systems to automatically adapt to previously unknown heterogeneous processors. This adaptive optimization avoids manual per-processor tuning. [http://cogadb.dfki.de/](http://cogadb.dfki.de/)

**PLASS** will develop a prototypical B2B platform for AI-based decision support for supply chain management. The focus is on the automatic recognition of decision-relevant information and the extraction of structured knowledge from global and multilingual text sources. PLASS enables SMEs and large enterprises to continuously monitor their suppliers and supply chains, and supports supply chain managers in risk assessment and decision-making. The DFKI plays a central role in the processing of unstructured, textual data. [https://www.plass.io/](https://www.plass.io/)
Current Projects

**Berlin Big Data Center Phase II (BBDC II)** will explore scalability issues surrounding the *real-time processing of data streams* and *declarative machine learning on massive datasets*. In addition, varying application areas will be addressed, including the analysis of distributed biomedical data and heterogeneous morphomolecular data arising in cancer research, learning on compressed data streams, real-time speech technology for interactive user interfaces, as well as security and privacy issues concerning the handling of sensitive personal information in Big Data systems.

[https://big-data-berlin.dima.tu-berlin.de/home/](https://big-data-berlin.dima.tu-berlin.de/home/)

**DAYSTREAM** (Data Analytcs and AI for Secure, Trusted, and REliAble Mobility) aims to develop data-driven applications for early detection, timely tracking and the best possible prediction of safety and process-critical mobility events


**ADAM** (approximate analysis of high-velocity data streams using modern hardware) will develop concepts for cost-efficient and scalable stream processing based on stream summaries and specialized hardware. This project is carried out jointly with Huawei Technologies.


In the **SIM3S** (Smart Inclusive Multi-Modal Mobility Services) project, data from the BMVI data offerings mCloud and MDM will be linked, refined and jointly analysed with other open data, user-generated content and data from individual modes of transport and other mobility-relevant companies in order to remove barriers and barriers to discrimination in everyday mobility. For the implementation of the project, state-of-the-art technologies and methods from the areas of Big Data Intelligent Analysis of mass data and artificial intelligence, in particular Natural Language Processing (NLP), are used.

ExDRa (Exploratory Data Science over Raw Data) aims to investigate suitable system support for the exploratory data science process on heterogeneous and distributed raw data sources and to provide it as part of a prototype for real-world use cases. In detail, the approach includes the following main research areas:

- Ad-hoc and federated data integration over raw data
- Data organization and reuse of intermediate results
- Horizontal optimizations across the entire data science life cycle
- Request planning for limited access to data


Lapse. is backed by Software Campus and aims to develop a system architecture that mitigates communication costs for distributed machine learning. To reduce communication, researchers have developed algorithms that exploit locality, whereby, workers solely update a subset of the model parameters, at a given time. Our intention is to yield a solution that is applicable to a wide-range of ML applications and aids in the development of advanced ML-based solutions for today’s societal challenges. This project is carried out jointly with Trumpf.

[https://www.dima.tu-berlin.de/menue/research/current_projects/lapse/](https://www.dima.tu-berlin.de/menue/research/current_projects/lapse/)

moreEVS (The Multi-Domain Modeling and Optimization of Integrated Renewable Energy and Urban Electric Vehicle Systems) aims to address the challenge of conducting large-scale data analysis efficiently and pair renewable energy power sources with electrical vehicles, to facilitate charging. The project enables state-of-the-art research to be conducted in energy demand modeling, battery quality management, grid infrastructure for renewable energy, and large-scale data management.

[https://www.dima.tu-berlin.de/menue/research/current_projects/moreevs/](https://www.dima.tu-berlin.de/menue/research/current_projects/moreevs/)
Current Projects

**Rhino** (Managing Very Large Distributed State for Scalable Stream Processing) will address scalable data stream processing and analytics challenges arising in Big Data, Cloud Computing, Industry 4.0, and IoT (Internet of Things) applications. Our aim is to develop a novel state management solution for scalable (i.e., low-latency, high-throughput) stream processing that enables fine-grained fault-tolerance, on-demand resource scaling, and load balancing in the presence of very large (e.g., hundreds of GBs) distributed state. We will devise a technological framework that seamlessly provides fault-tolerance, resource-scaling with zero downtime, and offers high-resource efficiency, lower operational costs, and reduced time-to-knowledge to end-users working on large-scale data applications. This project is part of Software Campus and is carried out jointly with Huawei.

[https://www.dima.tu-berlin.de/menue/research/current_projects/rhino/](https://www.dima.tu-berlin.de/menue/research/current_projects/rhino/)

**EDADS** (Efficient Data Analysis on Data Summaries) principally aims to design and implement sketch algorithms for streaming data in modern dataflow engines. Consequently, this would serve to reduce the size of data streams. Furthermore, in contrast to examining the entire dataset, the sketch could then be used by data analytics (e.g., for anomaly detection) and thereby shorten the data analysis execution time. This project is part of Software Campus and is carried out in cooperation with Holtzbrinck Publishing Group.

[https://www.dima.tu-berlin.de/menue/research/current_projects/edads/](https://www.dima.tu-berlin.de/menue/research/current_projects/edads/)

**BigMedilytics** aims to use state-of-the-art Big Data technologies to improve the productivity of the healthcare sector, by reducing patient costs, improve care quality via better patient outcomes, and deliver better access. *Our group* provides the Big Data Healthcare Analytics Blueprint and combines advanced diagnostic data from the Charité transplant center with data from ambulatory healthcare providers and real-time patient-level data transmitted via smartphone.

Completed Projects

**BIZWARE** developed a “model and software factory” and a runtime environment that enables the modeling, generation and execution of software components and applications based on domain-specific languages (DSLs). *Our group* led the “BIZWARE Factory” and worked in modeling, DSLs and modeling methodologies, information management in software engineering, software lifecycle management, semantic technologies as well as information extraction for software development.  
[https://www.dima.tu-berlin.de/menue/research/completed_projects/bizware/](https://www.dima.tu-berlin.de/menue/research/completed_projects/bizware/)

**BIZCYCLE** was to become the leading technology platform for cross-system, corporation-wide integration of software solutions and components for corporate management. *Our group* provided the software foundations and the methodology, with which the interoperability platform was being developed via the “Methods, Tools and Infrastructure for the components integration platform subproject of the Bizycle regional growth core (ISKI).”  
[http://www.bizycle.de/](http://www.bizycle.de/)

**eDV-Hera** (efficient data processing on heterogeneous hardware) developed methods and technologies to reduce the development effort required when writing data processing programs, that can fully utilize heterogeneous hardware architectures, and run efficiently. This project was carried out jointly with SAP.  
[https://www.dima.tu-berlin.de/menue/research/completed_projects/edv_hera/](https://www.dima.tu-berlin.de/menue/research/completed_projects/edv_hera/)

**SENSE** addresses data collection and analysis in the Internet of Things. It explores ways to dynamically tune data collection to current analyses. As a result, analysis systems gain control over the production of data streams. This allows for providing tailored data streams based on the data demand of applications, which enables them to connect to a vast amounts of sensors at small data transfer costs.  

**RADAR** demonstrated how to extract data from unstructured text drawn from the web using newly implemented data analytics, in order to facilitate decision-making.  
Completed Projects

**Go-OLAP** recognized users’ search intent and provided information in real-time from web data sources. The GOOLAP.info prototype addressed three critical problems. First, extracting entities and relations from a large corpus of news data. Second, recognizing users’ search intent. And lastly, displaying aggregated data extractions in a simple, fast and intuitive way.

http://www.goolap.info/

**MIA** (a cloud-based marketplace for information and analytics) developed a marketplace, which ensures the sustainable operation of a reliable and trusted platform for the production, provision and use of German Web data and other available data sources.

https://www.dima.tu-berlin.de/menue/research/completed_projects/mia/

**SINDPAD** developed an indexing and search technology for structured data sets, leveraged graphics adapters to support query execution and aimed to achieve unprecedented performance compared to conventional systems of equal cost. *Our group* played a significant role in the conceptual planning and implementation of algorithms for hybrid GPU/CPU processing. We analyzed query processing algorithms and devised metrics to compare the performance of GPU-operators and CPU-operators.

http://dima.tu-berlin.de/menue/research/completed_projects/sindpad_query_processing_on_gpus/

**Smart Data for Mobility** aims to develop a big data analytics platform for multi-modal smart mobility service. This platform is aimed to provide intelligent data management functions and services, which help the mobility service providers for optimal planning, predication and prognosis, and also support travellers for their individual travel planning.

https://www.sd4m.net

**Smart Data Web** aims to provide a bridge between publicly available information on the web and the internal information worlds of companies. By automatically combining public data streams, i.e., news websites or social media channels, with company internal data, production companies will be able to react faster and more securely to changes.

https://www.smartdataweb.de
Completed Projects

RETHINK big (a roadmap for European technologies in hardware and networking for big data) delivered a strategic roadmap for how technology advancements in hardware and networking can be exploited for big data analytics. Experts investigated how technical developments can be matched to advancements in applications, algorithms and systems.

http://rethinkbig-project.eu/

ROBUST developed methods to better comprehend and manage the business, social, and economic objectives of users, providers and hosts. The objectives included finding solutions for community risk management, large-scale data management, models of community polity and politics, community simulation and community data analysis.

Our group contributed to this project by developing scalable methods for processing huge amounts textual data as well as graph data representing the social structure of a community.

http://www.robust-project.eu/

The SCAPE project developed scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects.

http://scape-project.eu/

DOPA took data distribution to the next level, established a repository of valuable, distributed business data, benefitting data seekers and data providers. DOPA demonstrated the application of Stratosphere to a correlation analysis of microblogging and stock trade volume data using Meteor.

http://www.dopa-project.eu/
Completed Projects

**Stratosphere** conducted research in the areas of massively parallel data processing engines, a programming model for parallel data programming, robust optimization of declarative data flow programs, continuous re-optimization and adaptation of the execution, data cleansing, and text mining. *Our group* researched a programming model that extends a functional map/reduce programming model by adding second order functions.

http://stratosphere.eu/

**Myriad** is a toolkit for scalable parallel data generators that generates large synthetic data sets, according to a predefined schema and a set of statistical restrictions. It was also designed to tackle challenging problems, such as the benchmarking and testing of systems designed to handle web-scale amounts of data. It provides a fast and intuitive way to define custom data generators tailored towards the requirements of a concrete use-case.

http://www.dima.tu-berlin.de/menue/research/completed_projects/myriad/

**ELS** (an optimal energy control and failure management system) aimed to realize optimal energy efficiency for complex control systems used in railway systems. *Our group* played a significant role in the conceptualization of a knowledge database for relevant operational scenarios, identification and description of data streams, construction of efficient renewal strategies in the event of failures, and the articulation of functional and technical specifications.

http://www.dima.tu-berlin.de/menue/research/completed_projects/els/

**PROTEUS** investigated and developed ready-to-use scalable online machine learning algorithms and real-time interactive visual analytics to deal with extremely large data sets and data streams. *Our group* chaired the project’s scientific and technical committee and led the development of a high-throughput, low-latency software architecture for joint batch and stream processing. Apache Flink was extended to serve as basis for novel scalable online machine learning algorithms.

https://www.proteus-bigdata.com/
Completed Projects

**STREAMLINE** employs Apache Flink to jointly carry out data analysis for both *data at rest* and *data in motion* far more efficiently and at a much lower cost than with existing tools and technologies. *Our group* coordinates scientific and technical activities among its partners and performs technical research and development. The project will benefit from using Apache Flink as a well-established basis for developing a highly scalable, high-throughput real-time stream mining platform.

http://h2020-streamline-project.eu/

**SAGE** redefines data storage for the next decade, with the depth of capabilities for Exascale HPC alongside a breadth of future ‘Big Data’ applications. *Our group* integrated the advanced data analytics platform Apache Flink with a native object interface together with the data-local computations. Consequently, Apache Flink benefitted from the full performance and features offered by the storage platform, lifting its analytics processing capabilities to the exascale level.

http://www.sagestorage.eu

**SePiA.Pro** is developing an open and standardized service platform for the optimization of production chains across machinery producers. The optimization parameters are based on intelligent analysis and evaluation of sensor data and job parameters.

http://projekt-sepiapro.de/en/

**Smart Data Forum** connected users and developers of Smart Data technologies in Germany with the professional community both nationally and internationally and supported small and medium enterprises (SMEs) within their digitalization process. For this purpose, we provided an overview of the technology programs funded by the Federal Ministry of Economic Affairs and Energy (BMWi), offered guidance and education tailored to the needs of the SMEs and organized events with international stakeholders.

https://smartdataforum.de/en/
Completed Projects

Peel is a framework that helps you to define, execute, analyze, and share experiments for distributed systems and algorithms. A Peel package bundles together the configuration data, datasets, and workload applications required for the execution of a particular collection of experiments. Peel bundles can be largely decoupled from the underlying operational environment and easily migrated and reproduced to new environments.
http://peel-framework.org/

Berlin Big Data Center, Phase I (BBDC I) was established in 2014 and was a national Big Data competence center led by the Technische Universität Berlin (TUB). In Phase I, besides TUB, BBDC consortium partners included the Beuth University of Applied Sciences Berlin, the German Research Center for Artificial Intelligence (DFKI), the Fritz-Haber-Institute of the Max Planck Society, and the Zuse Institute Berlin (ZIB). Over its initial four-year period, the BBDC sought to prepare German/European industry, science and society for the global Big Data revolution. Its key objectives included conducting fundamental research to enable scalable big data analysis, developing an integrated, declarative, and highly scalable open-source system for advanced data analysis, transferring technology and know-how to support innovation in industry, and educating future data scientists at leading academic programs.
https://big-data-berlin.dima.tu-berlin.de/home/
Corporate Sponsors and Industrial Collaborators

 Corporate Sponsors

amazon, hp, HUAWEI, IBM, ORACLE, SAP, T-sys, zalando

 Industrial Collaborators

Alphabet, Audi, BOSCH, cisco, THALES, ZEISS, facebook, POLECAT, AWARE, SIEMENS, Model Labs, MESOSPHERE, Microsoft, neofox, ParStream, TERADATA, klopetek, Rolls-Royce, ClinPath, TTR, Okkam, Okumura, ISP, INDEO, modalplan GmbH, DB, BAHN, diamond, DAIMLER, Software Mind, adesso, FMK GmbH, software, klopetek, Cedavis, akquinet, modalplan GmbH, BAHN, TTR, Okkam, ISP, INDEO, INTEGRIS, ASCI, TEMIS, VICO, TRUMPF, holtbrueck, DATEV, Scheer, SAMSUNG, AMI, HFC, ArcelorMittal, TMT GmbH, Schiefer & innovative,いくらい!?®, jinit, S-Team, broov, MERCK, treelogic, KALEO, SEAGATE, OptiMedics, allinea, ROVIP, ARM, spark works, ubermetrics, EXUS, Atos, IMR, ida, lab.
Educational Programmes

- Pursue a Master’s or PhD with a focus on Information Management, Data Science, and Big Data.
- Learn, grow, and develop via our courses, research activities, and thesis opportunities.
- We enable students to develop fundamental skills relevant for today’s job market, entrepreneurs to start their own businesses, and prospective doctoral students to attain their PhD.
- Our curriculum is especially designed to ensure sound theoretical knowledge, supplemented with hands-on lab sessions, development projects, and seminars to deepen understanding.
- Guest lecturers from academia and industry often visit us to share their insights concerning current technologies and market trends.

**EU Sponsored Programmes**

**Member of Two Master’s Programmes**

- EIT Digital Master School Data Science (DSC) Major
- Erasmus Mundus - Big Data Management and Analytics (BDMA)
Representative List of TU Berlin DIMA Alumni

Prof. Dr. Alexander Löser
Full Professor, Beuth Hochschule
(previously, Sr. Researcher)

Prof. Dr. Asterios Katsifodimos
Assistant Professor, TU Delft
(previously, Postdoc)

Prof. Dr. Manohar Kaul
Assistant Professor, IIT Hyderabad
(previously, Postdoc)

Dr. Kostas Tzoumas
CEO/Co-founder, data Artisans
(previously, Postdoc)

Dr. Alan Akbik
Research Scientist, IBM Almaden
(previously, PhD Student)
Representative List of TU Berlin DIMA Alumni

Dr. Sebastian Schelter
Applied Scientist, Amazon
(previously, PhD Student)

Dr. Moritz Schubotz
Postdoc, Univ. of Konstanz
(previously, PhD Student)

Dr. Max Heimel
Software Engineer, Snowflake
(previously, PhD Student)

Dr. Stephan Ewen
CTO/Co-founder, data Artisans
(previously, PhD Student)

Dr. Fabian Hüske
Engineer/Co-founder, data Artisans
(previously, PhD Student)
Representative List of TU Berlin DIMA Alumni

Dr. Chen Xu
Associate Professor
East China Normal University (ECNU)
(previously, Postdoc)

Dr. Christoph Boden
IT Consultant
Mercedes Benz.io
(previously, PhD Student)

Dr. Alexander Alexandrov
Applied Scientist
Amazon
(previously, PhD Student)

Dr. Andreas Kunft
Chief Data Officer
Aignostics
(previously, PhD Student)